Document Title:	Recidivism Forecasting with Multi-Target Ensembles: Years One, Two and Three, Team TrueFit
Author(s):	David Lander, Russell D. Wolfinger
Document Number:	305048
Date Received:	July 2022
Award Number:	NIJ Recidivism Forecasting Challenge Winning Paper

Recidivism Forecasting with Multi-Target Ensembles

National Institute of Justice (NIJ) Recidivism Forecasting Challenge

Winning Solution for Male, Female, and Overall Categories in Year One, Team CrimeFree

- *David Lander, Northquay Capital, TrueFit.AI*
- *Russell D. Wolfinger, SAS Institute, Cary, NC*

Winning Solution for Male, Female, and Overall Categories in Year Two, Team TrueFit

- David Lander, Northquay Capital, TrueFit.AI

Winning Solution for Female, Runner-Up for Male and Overall in Year Three, Team TrueFit

- David Lander, Northquay Capital, TrueFit.AI

Abstract

Classification based on quantitative data is primarily about feature engineering and model ensembling. The former encodes and enriches patterns in the data, while the latter produces robust predictions even from a limited amount of data. Our winning solution for the Recidivism Forecasting Challenge included heavy amounts of each; we provide a roadmap to this solution, along with source code and guidance on how to produce similar results on future datasets.

Overview

The key to this competition was using all available information about each parolee, and using very large ensembles to stack these predictions forward. Models predicting other years, beyond each round's target, extract meaningful and relevant combinations of features that are also predictive of the target for each round. The overall ensemble for each round consisted primarily of gradient-boosted trees and neural networks. Massive ensembles of gradient-boosted trees, using wide and distinct sets of hyperparameters, were the most important models in predicting recidivism in years two and three.

Variables

Our models included all provided variables. The primary form of feature engineering was converting all textual features to ordered numerical categories, and grouping all PUMA regions by their regional composition. We also mapped future outcomes to a range of extra targets, such as a binary variable for each drug test category equaling 0%, 100%, or being unlisted. Full detail is available in the shared Github repository.

Models – Year Two and Three

In these rounds it was especially important to extract every bit of information about each feature. Gradient-boosted models, which split each feature into 255 bins, vastly outperformed neural networks or other "smooth" approaches; these models were likely most able to back into latent variables like the number of drug tests or total length of employment.

The final ensemble in each round included neural networks, with multi-path MLPs, and gradient-boosted trees, with a wide variety of hyperparameter settings. All models were fed into a linear model for averaging, in particular, an elastic net model where all weights must be greater than or equal to zero.

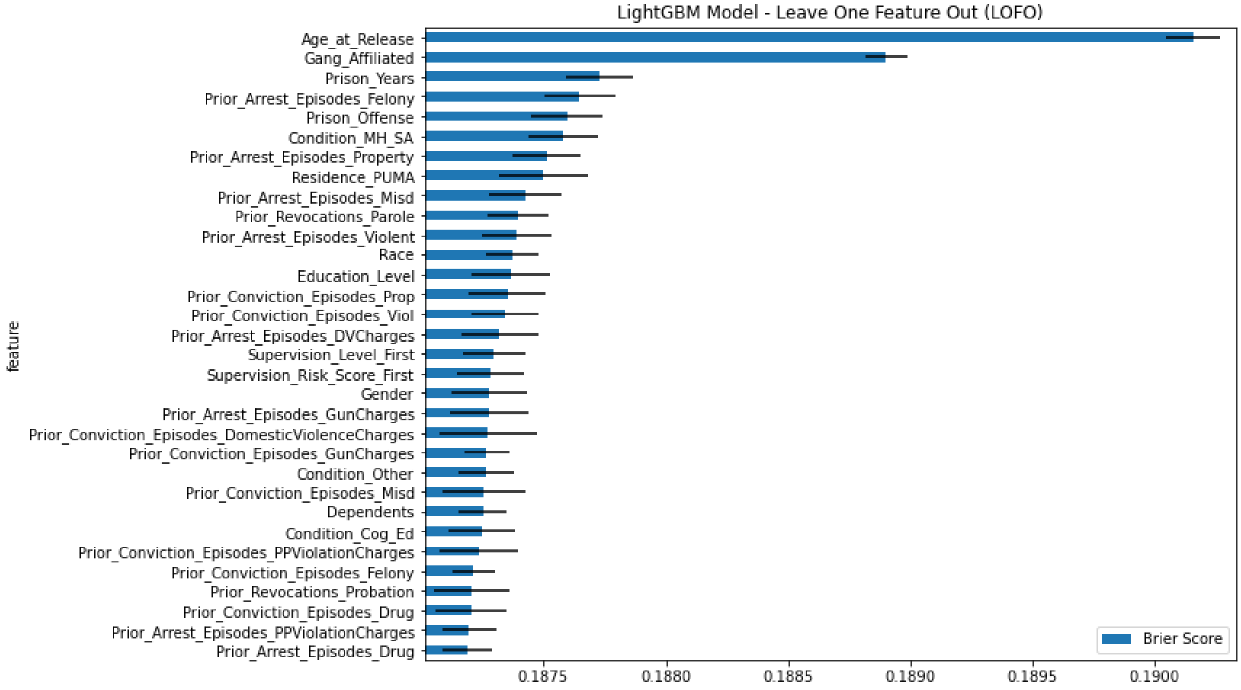
Cross-Validation

Models were generally trained on 80% of the data, while using the remaining data to estimate model performance. Given how easy it is to overfit any one model to any portion of the data, our process typically included training around 50-100 ‘folds’ of any given model, so that any data point was estimated as the average of at least 10-20 different model predictions.

Feature Importance

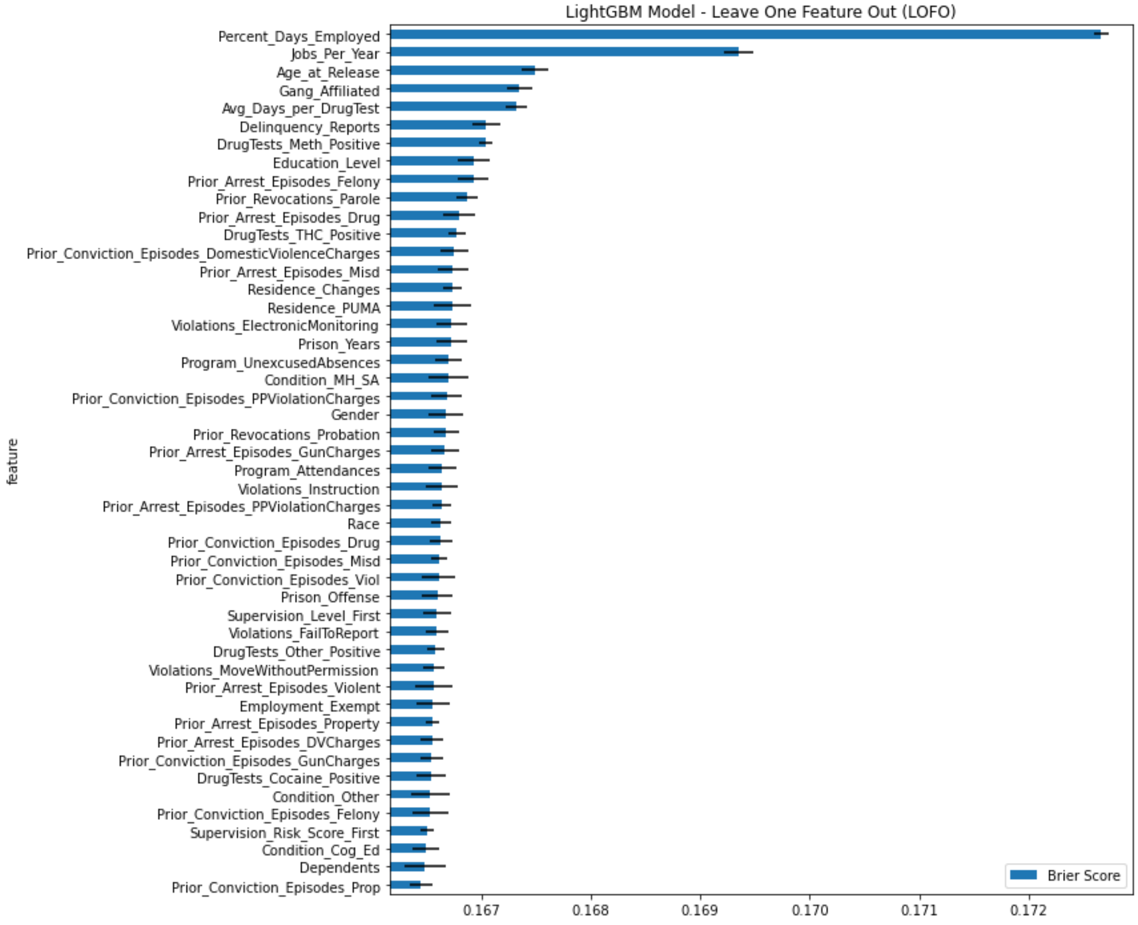
Estimating the importance of every feature is possible by producing models that “leave one feature out.” This analysis is shown below, along with confidence intervals to determine the statistical significance of each feature; features listed in the table below produce a statistically significance change in performance when left out.

Feature Importance -- Year One



Feature Leave-Out	Brier Score
Age_at_Release	0.1897
Gang_Affiliated	0.1883
Prison_Years	0.1873
Condition_MH_SA	0.1872
Prior_Arrest_Episodes_Felony	0.1872
Residence_PUMA	0.1871
Education_Level	0.1870
Prison_Offense	0.1870
Prior_Arrest_Episodes_Property	0.1870
Prior_Revocations_Parole	0.1870
Gender	0.1869
Prior_Arrest_Episodes_Misd	0.1869
Prior_Arrest_Episodes_PPViolationCharges	0.1869
Prior_Arrest_Episodes_Violent	0.1869

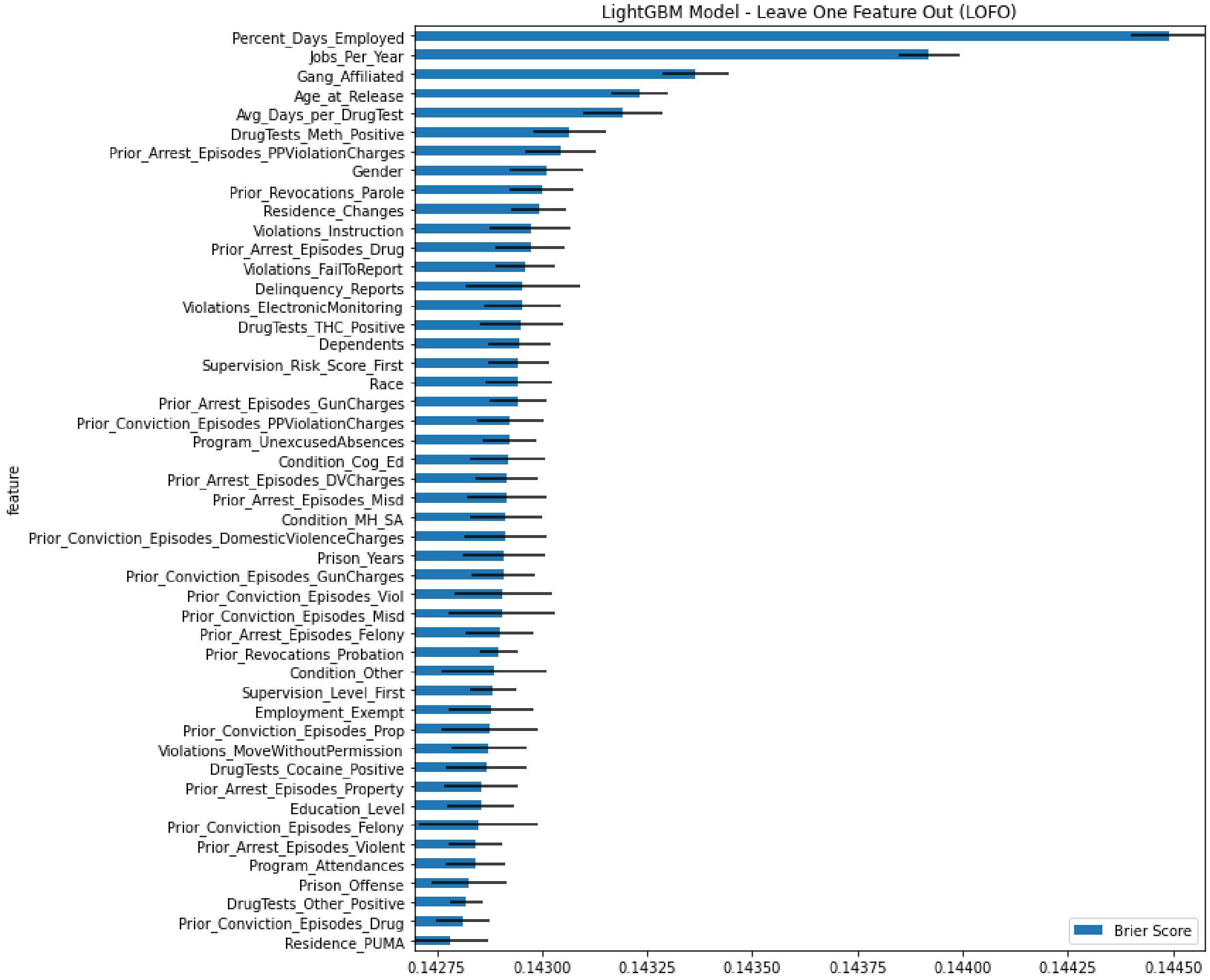
Feature Importance -- Year Two



Feature Leave-Out	Brier Score
Percent_Days_Employed	0.1727
Jobs_Per_Year	0.1694
Age_at_Release	0.1675
Gang_Affiliated	0.1674
Avg_Days_per_DrugTest	0.1673
Delinquency_Reports	0.1670
DrugTests_Meth_Positive	0.1670
Education_Level	0.1669
Prior_Arrest_Episodes_Felony	0.1669
Prior_Revocations_Parole	0.1669
Prior_Arrest_Episodes_Drug	0.1668
DrugTests_THC_Positive	0.1668
Residence_Changes	0.1667

This resource was prepared by the author(s) using Federal funds provided by the U.S.
Department of Justice. Opinions or points of view expressed are those of the author(s) and do not
necessarily reflect the official position or policies of the U.S. Department of Justice.

Feature Importance -- Year Three



	Brier Score
Feature	
Percent_Days_Employed	0.1445
Jobs_Per_Year	0.1439
Gang_Affiliated	0.1434
Age_at_Release	0.1432
Avg_Days_per_DrugTest	0.1432
DrugTests_Meth_Positive	0.1431
Prior_Arrest_Episodes_PPViolationCharges	0.1430
Gender	0.1430
Residence_Changes	0.1430

This resource was prepared by the author(s) using Federal funds provided by the U.S. Department of Justice. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

Feature Importance – Discussion

The primary takeaway from these plots of feature importance is that a few key features produce enormous changes in performance. In particular:

- Knowing the percent of days employed post-release, and number of jobs per year, is the single most important feature category
- Age at Release is important across all models: we observed that younger parolees are more likely to recidivate
- Gang affiliation was a key feature in all years
- Days per Drug Test and methamphetamine drug tests were the most predictive drug-test-related features

Accounting for Racial Bias

All of our solutions were tuned to optimize the overall metric for the competition. This approach incidentally scored 1st and 3rd in two competitions that accounted for racial disparities.

One common objection to 'black box' models is they may be particularly biased—even in cases where those biases help make accurate predictions; our experience here was that a winning solution tuned for overall accuracy happened to have minimal bias on a well-designed measure of racial bias. In addition, the 'Race' variable was not essential to forecasting, and did not make the list of statistically significant features in any round. Both of these findings indicate that models that do not consider race perform very well, and even those that include race as a variable do not exhibit material amounts of bias in forecasting recidivism.

Source Code

A full solution capable of producing strong results in all rounds of the competition is open-sourced at https://github.com/david-1013/RFC. This source code can be trained on a single machine within a day, and includes the full stack from data intake to model ensembling. This code is MIT-licensed, with no restrictions on its use for any purpose.

The included feature engineering code provides a roadmap for converting categorical features into ordered categories, and into one-hot-encoded variables for linear models and neural networks. The modeling code is self-tuning, and capable of producing strong solutions even with additional features or a completely different dataset. Finally, the ensemble code will select the best overall blend of models, and should be reasonably robust to domain-shifts, e.g. to other geographies.

Future Considerations

We found the competition to be well-structured with more than sufficient data to yield highly informative predictions. The final two rounds may have included 'beyond the decimal point' precision in some items that provided models with some clue of when recidivism occurred (e.g. values indicating a large number of drug tests would show a parolee remained crime-free for longer).

One interesting variation on the competition would be predicting how soon a parolee committed a crime, e.g. the target would be 1/sqrt(months to recidivism), equating to zero if no arrest occurred, thus indicating risk at release across all future time periods.

Conclusion

We thank the organizers for an interesting series of challenges. Our main go-forward suggestion would be providing real-world deployed models with as much information as possible. High-powered machine learning techniques are capable of extracting enormous amounts of signal from every possible feature, and even across numerous related domains. Age, employment, and gang affiliation were key drivers of model performance, and any additional related features would almost certainly produce further gains in predictive performance.

Machine learning models are clearly up to the task of predicting recidivism for the purposes of parole judgments or monitoring—the predicted probabilities spanned the full range from near-zero to very likely—and we look forward to future competitions and seeing evidence-based methods like this put into practice.